CULTURAL BIAS IN LANGUAGE TESTING

Patrisius Istiarto Djiwandono
Pusat Bahasa Universitas Surabaya, Surabaya

Abstract: An issue that has recently been gaining more attention in the domain of language testing is item bias. Defined as the characteristic of an item which causes learners of the same abilities but of different social groups to perform differently, the bias can be present as gender, ethnic, religious, social class, or cultural bias. The paper brings up a discussion in this area of concern by starting off from the concept of culture. It then explains what cultural bias is, how it manifests in items of language tests, and what unfavorable results may happen. It then highlights some examples of cultural bias in some popular standardized tests constructed by both foreign institutions and domestic educational bodies. The paper argues that if language teaching is to foster the development of cross-cultural understanding, attempts should be made to eliminate such bias. It then concludes by discussing Differential Item Functioning, a method that is used for dealing with bias in language tests.

Key words: language testing, cultural bias, item bias, Differential Item Functioning.

Learning English involves learning its culture. When learning how to express ideas in the language, learners simultaneously learn how to think in the same way an English native speaker thinks. The learner's awareness of the different cultural manifestations of the target language is encouraged, and is indeed a positive sign which indicates the learner's integrative motivation. Nevertheless, in the domain of language testing, the issue of cultural differences must be treated seriously if the test is to treat all test-takers fairly. The paper addresses the issue of cultural bias in language testing. It starts with a definition of culture and cultural bias, then proceeds to demonstrate how the bias manifests in some language tests. Finally, it concludes by touching upon Differential Item Functioning, a statistical analysis designed to detect and avoid such bias in a language test.

CULTURE DEFINED

The current literature abounds with definitions of culture. The most recent definition of culture is given by Harris (1999), who asserts that a culture is a set of values, beliefs, and norms that members of a particular community acquire and use in interacting with the world. Leininger (1991) provides a more concise definition for the concept of culture as the "learned, shared and transmitted values, beliefs, norms, and lifeways of a particular group that guides their thinking, decisions, and actions in patterned ways." As such, culture covers all aspects of social life, including both thought and behavior. Two of the many manifestations of culture are symbols and values. Symbols, which also include words, represent the outermost layer of a culture, while values are its deepest manifestation. For instance, the pronoun and honorific system of English language manifest an egalitarian culture, and as such stands in stark contrast with those of Javanese language, which reflect a more stratified society.

Because of the existence of many cultures in the world, one can speak of cultural differences and cultural relativism. The latter means that there is no universal standard for the most acceptable behavior and attitudes. Judgment concerning the acceptability of particular behavior or attitude depends on one's cultural point of view. In a test of spoken abilities, a Western examiner who is not aware of this notion may interpret an Asian test-taker's little verbal responses as a sign of lacking verbal abilities, not realizing that the Asian examinee is just being polite by not saying too much. If cross-cultural awareness is not promoted among test writers, examiners, raters, and interviewers, such misunderstanding may disadvantage learners who are otherwise quite proficient in their spoken English.

TEST FAIRNESS

A test is said to be fair if the "construct-irrelevant personal characteristics of test-takers have no appreciable effect on test results of their interpretation" (Educational Testing Service, 2000:17). In other words, a test is fair if test-takers with the same ability or equal standing on the construct being measured earn the same score irrespective of their native cultures. A test which is designed to measure grammatical ability, for instance, should not contain items which measure mathematical ability because the latter skill is not relevant to the construct being measured, i.e. grammatical competence. Such items will cause the scores to vary, and the variance surely cannot be attributed to the learners' grammatical ability. Learners who happen to have higher mathematical skills are clearly advantaged over the less able learners,

which is an indication that the items have been unfair to them. In other words, the items are biased against those having low mathematical ability. By the same token, items designed to measure reading abilities in English should not measure knowledge about a particular custom in Western world that only an English native speaker is familiar with. If that is the case, the items are said to be culturally biased against learners of non-Western cultures. In both cases above, the items are unfair to some of the test-takers.

Test unfairness may be induced by test bias, which, according to Brown (2004:111) can be in the form of language, culture, race, gender, and learning styles. With respect to culture, care must be taken not to introduce elements in the test that disadvantage learners of different native cultures.

EXAMPLES OF CULTURALLY-BIASED ITEMS

When discussing cultural aspect in English language testing, it is customary to refer to the non-Western learners as the focal group, and the English native speakers as the reference group. The two groups may be nurtured by widely contrasting cultures. Take for example the contrast between Indonesian and American culture. According to Draine and Hall (1986), Indonesian learners are brought up in a culture which puts special importance on status. In this culture, people are of unequal status, and therefore the concept of respecting others in accordance to their status must be exercised appropriately. This is even more evident in Javanese culture, whose language consists of levels that must be used in accordance with a speaker's and her interlocutor's rank, status, age, and degree of acquaintance. American culture, in contrast, is characterized by individualism and self reliance. Americans are more spontaneous and informal, but also confident, logical, direct, passionate about truth, justice and equal opportunity (Millet, 2000).

When writing test items, English native speakers may incidentally include items that are manifestation of their native culture. Items of this kind potentially confound non-Western examinees because of their cultural load, causing the examinees to fail to answer them correctly. McGinley (2002) points out that a few standardized tests used in the US contain items that may be considered culturally bias. She mentions Woodcock-Johnson Revised, a test which is replete with items describing nursery rhyme and American pop culture, which are probably unfamiliar to EFL learners from some other cultures.

Other real examples of this potential bias are evident in the following items, taken from TOEFL Test Preparation Kit Workbook, published by Educational Testing Service (1998).

A listening comprehension item number 29 on page 280 presents the following dialogue:

> (man) I'm taking up a collection for the jazz band. Would you like to give?
> (woman) Just a minute while I get my wallet
> (narrator) What will the woman probably do next?
>
> 29. a. put some money in her wallet
> b. buy a band-concert ticket
> c. make a donation
> d. lend the man some money

The right answer is c. However, it is very unlikely that examinees of non-Western culture are familiar with the habit of collecting money for a band in the USA culture. Therefore, being largely unfamiliar with the meaning of "taking up a collection" and looking at the word "give" from the man, they may be misled into thinking that the answer is d. Alternatively, if they have no idea whatsoever that a band in the US may need to collect some money, they may choose b.

Another listening comprehension item, number 16 on page 290, gives this conversation:

> (man) Can you go over my notes with me? I'll never understand all these chemistry experiments.
> (woman) You know, review sessions are being held every night this week. They are supposed to be good.
> (narrator) What does the woman imply the man should do?
>
> 16. a. make a copy of his notes for her
> b. ask his professor for help
> c. attend the review sessions
> d. go to the chemistry lab this evening

Examinees who come from an academic culture in which universities normally do not hold review sessions after classes may be at a loss to figure out the right answer to the item above. At least, being unfamiliar to the habit

of review session on US campuses, they may have to spend a little more time than necessary to get at the right answer, which is c.

In the reading comprehension section, after a passage about the development of US politics, number 6 on page 28 is given as follows:

> It can be inferred from the passage that early hotelkeepers in the US were
> a. active politicians
> b. European immigrants
> c. professional builders
> d. influential citizens

If the examinees have only vague knowledge about the historical development of the US, the reading passage will not help them make the right inference about the early hotelkeepers in the US. As a result, it is likely that they will have problem selecting the right option. Thus, this inability to answer correctly is not attributable to their reading ability but to their unfamiliarity with the cultural aspect in the item.

An even more obvious example is shown by the item below, which appeared in the State Examination of English Language for Elementary Level (Ujian Nasional Diklusemas Bahasa Inggris Dasar Satu) on 26 Agustus 2001:

> 46. Mia: What should I do with this "*martabak*?"
> Mom: Just put them on a (a) drawer, (b) plate, (c) stove, (d) mug

Examinees that come from Java or are somehow familiar to Indian culture will find the item above easy to answer. Yet, for some others coming from different areas, the word "*martabak*" may be entirely new and therefore they do not know whether a "*martabak*" refers to a kind of stationery, food, a cooking device, or a kind of beverage. Despite their excellent English proficiency, there is no way they can get at the right answer. The item, in other words, is culturally biased against these examinees.

HOW TO DETECT THE PRESENCE OF CULTURAL BIAS IN A LANGUAGE TEST

A test given to a group of test-takers who are of different cultures, ethnicities, and gender can be considered fair if it exhibits measurement equivalence. Drasgow (1984:134) states that measurement equivalence exists "when the relations between observed test scores and the latent attribute measured by a test are identical across subpopulations". In simpler words, a test is fair if the scores reflect the test-takers' varying levels of proficiency in the skill that the test measures, and not their differences in any other aspect. A test which aims to measure listening comprehension, for instance, may not be fair--and therefore biased--if the scores reflect the differences among the test-takers in terms of their familiarity with a particular culture. Test-takers not familiar with the culture may score low not because they have poor listening skill but because they lack knowledge about the cultural aspect that somehow "sneaked" into the items. A rather extreme example is exhibited by an item in a listening comprehension test that goes as follows:

> Voice from the tape:
> *Camping is a popular form of recreation in the United States. When going camping, people bring a tent and usually a cooking stove. The campers cook in turn. Although women usually cook, men often do the cooking, too.*
>
> Question: *From the information about who cooks the meal during camping, what can you infer about the role of women and men in doing household tasks in the US?*

The success in answering the above item depends very much on a test-taker's familiarity with the nature of relationships between men and women in American culture. The test-taker may catch the spoken ideas very well, but if she knows vaguely about the men and women relationship in the US, she may fail to answer the item. Conversely, if another test taker happens to have lots of US friends and therefore knows much about the gender issue in the country, she will be able to answer the item. Thus, the scores of the two test-takers for this item will be different not because they have different listening skill, but because of the difference in their knowledge about the culture of the target language. The item, in other words, is culturally biased against the first test-taker.

To date, the presence of biased items is detectable through a measurement model which is commonly called Item Response Theory (IRT). Davies

et al. (1999:98) defines IRT as a body of theory that aims to infer from test scores highly stable estimates of test-takers' abilities and item characteristics. Compared to the traditional technique of determining test takers' abilities and the properties of items (their level of difficulty and their level of discrimination) which are somehow influenced by the nature of the test-takers who took the test, IRT generates a much more stable predictions across a much wider population of test-takers. From these stable estimates, a number of advantages can be derived, one of which is the ability to detect biased items such as the one exemplified above. Henning (1987:114) sheds light on the advantage of using IRT to deal with item bias by offering this more straightforward explanation:

> Item Response Theory has the advantage that it permits the quantification of the magnitude and direction of bias for individual items or persons. This enables the correction of test bias whether through removal, revision, or counterbalancing of biased items.

Once fed into the IRT analysis, the data from the test scores are expressed as differences on the delta scale (ETS, 2003), a scale that ETS created to place test items on a ranking of difficulty. A negative value indicates that an item is more difficult for the focal group, while a positive value means that it is more challenging to the reference group. The higher the number, the higher the difference between the two groups.

Thus, a test item like the above clearly fails to follow the principle of measurement equivalence. When analyzed using Item Response Theory model (IRT), such item will indicate differential item functioning (DIF), defined rather formally as "a statistical property of a test item in which different groups of test takers who have equal ability in the construct being measured have different average item scores because they are of different sociocultural groups".

An example will clarify the concept of DIF with reference to the procedure used by ETS above. Suppose a multiple-choice test of Reading Comprehension is given to a group of multicultural students who have been matched according to their reading abilities. On one particular item, say, item number 16, the American students choose the correct option "B" 15 percent more often than did the Asian students, who choose the incorrect options "D" and "A". This may be an indication of DIF, though there is no definite certainty about this unless the delta scale procedure is applied as follows. First of all, a preliminary step called Mantel-Haenszel odds ratio calculates the chances that the reference group (the Americans) will get the item correct, and that the focal group (the Asians) will get it correct. Then, it

divides reference group chances by focal group chances to show the possibility for the reference group to get the item correct. The ETS delta scale procedure then classifies the item as one of three types: (1) negligible DIF, where the resulting figure, called the effect size, is less than one delta unit, (2) intermediate DIF, and (3) large DIF, where the effect size exceeds 1.5 delta units and is significantly larger than 1 delta unit. If item number 16 falls into type (3), it can be concluded that though the reference group (American students) and the focal group (Asian students) are on equal reading comprehension ability, their average scores for the item differ significantly. A closer scrutiny of the item in question may reveal that it tests the knowledge of American culture more than it measures reading comprehension. It should also be clear from the above example that the item treats the test-takers unfairly by boosting the scores of the American students and lowering those of the Asian students. The item clearly has DIF, is unfair and thus needs to be replaced.

CONCLUSION

The paper addresses the issue of cultural bias in language testing. A test item is said to be culturally biased if learners from certain sociocultural groups fail to answer the item correctly because they do not have sufficient knowledge about the culture of the target language. Some real items which are very likely to contain cultural bias are presented. In keeping up with the spirit of cross cultural understanding, attempts must be made to weed out culturally biased items from high stake tests so as to maintain a fair assessment of language proficiency across learners of different cultures. One of the ways to overcome cultural bias is by submitting the scores of a test to an Item Response Theory analysis, and to detect whether any particular item shows differential item functioning (DIF). The presence of DIF in an item indicates that learners of equal ability on the skill being tested have performed differently because they differ in terms of some cultural aspects. Items with DIF are omitted from the scoring, and, if the test is subject to further revision, the items will be replaced with non-biased items.

REFERENCES

Brown, H. D. 2004. *Language Assessment: Principles and Classroom Practices*. NY: Pearson Education.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., and McNamara, T. 1999. *Dictionary of Language Testing*. Cambridge: University of Cambridge

Draine, C., and Hall, B. 1986. *Culture Shock.* Singapore: Times Book International.

Drasgow, F. 1984. Scrutinizing Psychological Tests: Measurement Equivalence and Equivalent Relations with External Variables are the Central Issues. *Psychological Bulletin*, 95, pp. 134-135.

Educational Testing Service. 1998. *TOEFL Test Preparation Kit Workbook. Princeton*, New Jersey.

Educational Testing Service. 2000. *ETS Standards for Quality and Fairness.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. 2003. What's the DIF? Helping to Ensure Test Question Fairness. Retrieved 25 March 2006, from http:// www.ets.org/research/dif.html

Harris, M. 1999. *Theories of Culture in Postmodern Times.* Walnut Greek, CA: AltaMira Press.

Henning, G. 1987. *A Guide to Language Testing: Development, Evaluation, Research.* Boston, Massachussetts: Heinle & Heinle Publishers.

Leininger, M. M. 1991. *Culture Care Diversity and Universality: A Theory of Nursing.* New York: National League of Nursing.

McGinley, S. 2002. Standardized Testing and Cultural Bias. *ESOL Multicultural Newsletter*, December 2002. Kansas: Fort Hayes State University.

Millet, J. 2000. *Understanding American Culture: from Melting Pot to Salad Bowl.* http://www.asia links.com/biz/sv/csarticle.asp?articleid =12398, 16 October 2004.